RGBD object recognition and visual texture classification for indoor semantic mapping

David Filliat, Emmanuel Battesti, Stéphane Bazeille, Guillaume Duceux, Alexander Gepperth, Lotfi Harrath, Islem Jebari, Rafael Pereira, Adriana Tapus, Cedric Meyer, et al.

RGBD object recognition and visual texture classification for indoor semantic mapping

David Filliat, Emmanuel Battesti, Stéphane Bazeille
Guillaume Duceux, Alexander Gepperth, Lotfi Harrath
Islem Jebari, Rafael Pereira, Adriana Tapus

Unite Electronique et Informatique
ENSTA ParisTech
32 bvd Victor, 75015 Paris, France

Email : david.filliat@ensta-paristech.fr

Cedric Meyer, Sio-Hoi Ieng, Ryad Benosman
Institut De la Vision
Université Pierre et Marie Curie
4 place Jussieu, 75005 Paris, France

Eddy Cizeron, Jean-Charles Mamanna, Benoit Pothier
GOSTAI S.A.S
15 rue J.B. Berlier, 75013 Paris, France

***Abstract*—We present a mobile robot whose goal is to autonomously explore an unknown indoor environment and to build a semantic map containing high-level information similar to those extracted by humans. This information includes the rooms, their connectivity, the objects they contain and the material of the walls and ground. This robot was developed in order to participate in a French exploration and mapping contest called CAROTTE whose goal is to produce easily interpretable maps of an unknown environment. In particular we present our object detection approach based on a color+depth camera that fuse 3D, color and texture information through a neural network for robust object recognition. We also present the material recognition approach based on machine learning applied to vision. We demonstrate the performances of these modules on image databases and provide examples on the full system working in real environments.**

I. Introduction

The problem of Simultaneous Localization and Mapping (SLAM) of an unknown environment by a mobile robot has been the subject of intense research for more than 20 years. Today, very robust solutions exist for SLAM in planar environments using 2D scanning laser sensors to the point where several effective commercial or open source software packages are available. Research on SLAM is now more focused on 3D SLAM using range sensing, vision or a combination of these two modalities, in particular thanks to the recent development of low cost depth sensing cameras. Fusing this information with color imaging leads to RGBD sensors that have a huge potential for SLAM, object recognition and human-robot interaction.

Beside approaches that are mainly directed toward low level robot localization and mapping, several approaches have been proposed to introduce higher-level semantic information into maps. This includes the classification of space into different categories such as rooms, corridors [1], roads, buildings [2] and the addition of objects in a hierarchical map representation. For example, [3] propose an approach to object search that takes advantage of the organization of space into rooms, furnitures and objects laid in or on the furnitures.

This paper presents a mobile robot whose goal is to perform autonomous semantic mapping of indoor environments. This robot has been developed during the PACOM[1] project in order to participate in the "CAROTTE" challenge organised by the french Armament Procurement Agency (DGA) and Research Funding Agency (ANR). This challenge proceeds over three years with an increase in the difficulty over the years. The competition between 5 selected teams takes place in an arena of approximately 120 m^2 wherein objects and obstacles are placed (Figure 1). The environment contains several rooms, with variable ground types and difficulties (fitted carpet, tiling, grid, gravel, ...). Several types of objects are present in multiple instances, either isolated or in groups, which must be detected, located, and identified or characterized by the robot. Chairs, computers, boxes, books and plants are examples of the objects used in the competition. The complete description of the environment can be found on the challenge website[2].

Fig. 1. View of our robot in the arena during the CAROTTE competition.

The robot described in this paper has participated in the second CAROTTE competition and is an evolution of our previous work [4]. Most notably, we have modified the robot structure and included a RGBD camera that is used mainly

[1]http://cogrob.ensta-paristech.fr/pacom/
[2]http://www.defi-carotte.fr

for object detection. We also added the capability to recognize the ground and wall material using vision and improved the obstacle detection in order to deal with 3D obstacles and gravel areas that our robot cannot cross.

The remainder of the paper presents a short state of the art on object and texture recognition, before presenting an overview of the robot architecture and detailing our approaches to object recognition, ground and wall analysis and robust multi-cue obstacle avoidance. We finally present the semantic maps built by our system before analyzing its current shortcomings and its future improvements.

II. Related work

Visual object recognition is traditionally conducted in the context of two-dimensional images [5], [6], [7], [8], [9]. With the recent advent of off-the-shelf 3D sensors the recognition based on 3D information has flourished considerably [10], [11], [12]. In the 2D domain, several popular methods ignore geometric information altogether, using mainly histograms over colors [5] or significant sub-patterns [8], [13], whereas others emphasize hierarchical geometric relations [14] or explicit object parts [15]. Some geometry-based approaches such as histograms-of-gradients [16], [17] are currently very popular since they can be used in real-time on standard hardware. SIFT or SURF descriptors [6], [7] are often used to approximate histograms of gradients since their execution speed is even higher. Object recognition based on 3D features and descriptors is a relatively new but fast-growing area [10], [11], [12] where typical problems of 2D recognition, such as rotation and translation invariance, are less restrictive.

Ground and wall type classification is required in the competition and is also necessary to avoid gravel areas that our robot cannot cross. We chose to use vision for this task, reducing the problem to that of visual texture classification. Many methods for texture classification have been proposed relying on filter banks. These filters encode the local spatial variation that characterize a texture (e.g., [18], [19]) and are used as pre-processing for a classification algorithm such as Support Vector Machine of Nearest Neighbour. However computing such filter may be computationally expensive and other approaches have been shown to give better performances by directly processing image patches [20]. In particular, the approach proposed by [21] relies on randomized trees applied directly to random image sub-windows without any pre-processing and provides very good performances at a small computational cost.

III. System Overview

A. System Architecture

Our robot (Figure 2) is based on a pioneer 3 dx from Mobile Robots Inc. The robot was fitted with a SICK laser range finder, a ring of sonar sensors, a Pan-Tilt-Zoom color camera and a Microsoft Kinect camera as RGBD sensor. Three on-board computers linked through an Ethernet router run all the software modules involved in semantic mapping.

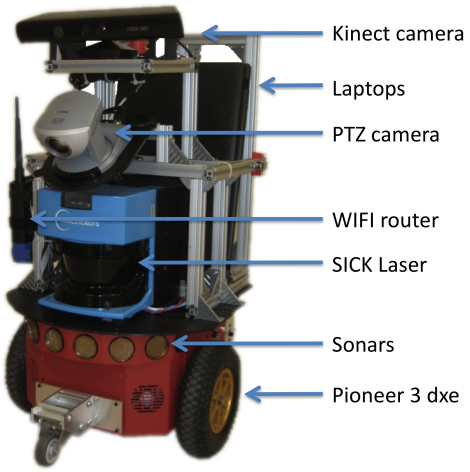

Fig. 2. The robot developed for the PACOM project.

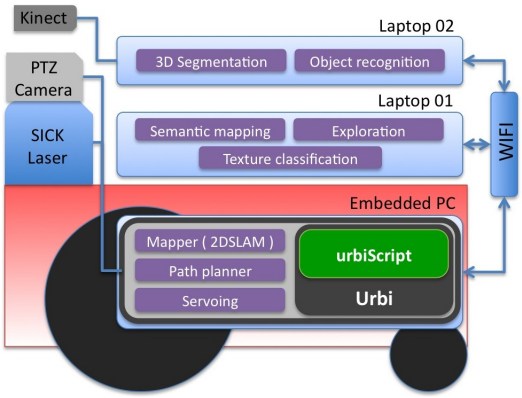

Fig. 3. The software architecture of our robot showing the repartition of the software modules on the 3 onboard computers.

The software architecture uses the Urbi framework[3]; an open-source middleware for programming complex robotic systems developed and supported by Gostai. Urbi is composed of a distributed component architecture (UObject), and an innovative orchestrator language (urbiScript) to coordinate all components. This language incorporates high-level features that facilitate the development of parallel and event-based applications. For the project, we thus developed a set of UObjects in C++ carrying out the various necessary functionalities (Figure 3). The whole mission of the robot is implemented in urbiScript which uses these UObjects' functionalities and coordinates their activation.

The 2D mapper using the SICK laser range finder and path planning UObjects are based on the Karto software library from SRI International[4]. The robot is controlled via the servoing UObject that implements a simple PID controller for trajectory following. The multi-objective exploration algorithm (described in detail in [22]) is in charge of choosing exploration points in order to discover the whole environment

[3]http://www.urbiforge.org/
[4]http://www.kartorobotics.com/

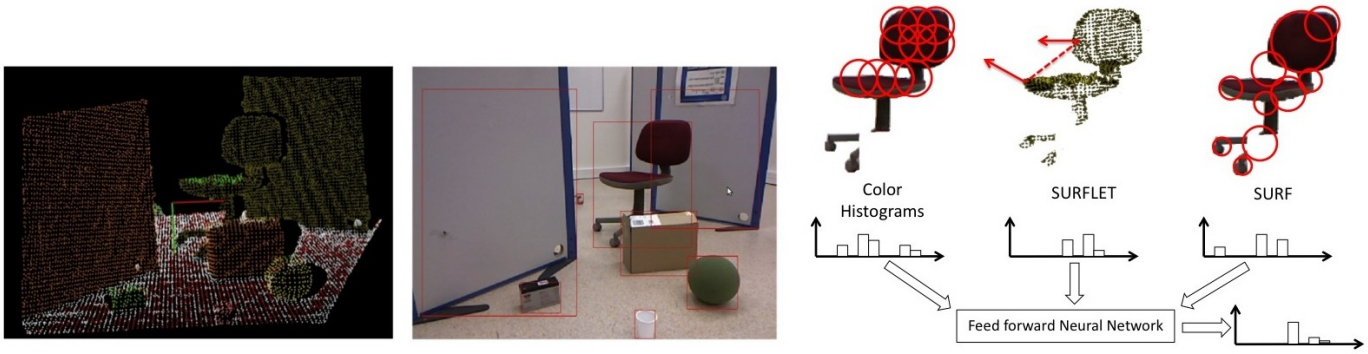

Fig. 4. Object detection and recognition process overview. Left: detection of floor points (shown in red) in the 3D sensor data, followed by segmentation of object clusters indicated by different point colors. Middle: 2D object bounding boxes computed by 3D to 2D projection of object clusters. Right: composite object detection method on the basis of the 3D object clusters and 2D object masks.

through the laser sensor and to search for objects everywhere in known space through the Kinect camera. The object recognition, texture classification and semantic mapping modules will be described in the next sections.

B. Object detection

Object detection in our system is realized by several processing steps making use of both RGB and depth information: floor detection, object segmentation and object recognition, see Fig. 4. We make use of the PCL library for 3D point clouds processing [23].

a) Floor detection and removal: Detection and subsequent removal of floor points is performed on point cloud data obtained from the Kinect sensor. This is an essential step as the floor prevents correct segmentation by connecting all objects standing on it. As a full RANSAC step for detecting planar surfaces would be too time-consuming for live operation, a simpler approach is applied which makes strong use of heuristics yet is able to localize floor points with great reliability. To this end, an initial calibration step conducts a full RANSAC analysis to detect planar surfaces in a test setting where a large floor area is visible. The model parameters of the largest planar surface are retained as the ”theoretical floor plane”. In the course of online processing, every new point cloud obtained from the sensor is downsampled (”voxelized” with a 2 cm step) to reduce noise, and local surface normals are computed. To identify floor points, we impose a constraint of elevation (no more than 20 cm over the theoretical floor plane) and normal (no more than 4 degrees deviation from the theoretical floor normal). If a sufficient number of points is found in this way, the ”actual floor plane” model is then estimated from them, otherwise the actual floor plane model is set to the theoretical floor plane parameters. The latter case can occur if the floor is not visible due to occlusions or adjacent walls. Finally, all 3D points are removed whose distance to the actual floor plane is smaller than 2 cm, or larger than 2 m in order to remove elements perceived outside of the competition arena.

b) Point cloud segmentation: After the removal of the floor, the remaining 3D points can easily be grouped into disjunct segments by a simple volume growing process of disjunct clusters, adding new points to clusters if their distance to the closest cluster points is less than 4 cm. As a post-processing step, we merge clusters if they contain points less than 4 cm distant when projected on the actual floor plane. This step is performed because there are complex objects such as chairs that get segmented into several clusters one above another when the object parts connecting the clusters are occluded. Each of the final point cloud segments is assumed to correspond to an object, and a tight 2D mask is computed for all objects by reprojection of 3D points to the image plane. These masks are an important prerequisite for efficient appearance-based object recognition based on the RGB image.

c) Wall detection: Each segmented 3D point cluster is checked for potentially being a wall using heuristics similar to the identification of floor points: if a sufficient percentage of points (90%) has a surface normal perpendicular to that of the actual floor plane, we assume this cluster is actually part of a wall. It is important to note that this method does not always work reliably, since large objects such as cupboards fulfil similar conditions, thus leading to missed detections in certain viewing conditions of these objects. The detected walls are subsequently categorized according to the method described in section III-C.

d) Object recognition: The remaining object clusters should then be identified. In order to obtain robust recognition, we combined three state-of-the-art recognition methods using a feed-forward neural network. Individual methods are local color histograms [24], SURF keypoints [7] and 3D surflet [10]. Especially the last method is novel for our system because it is intrinsically based on point cloud data. The two other methods have been used in our previous system [4], but are applied here only to the area resulting of the 3D points projection, thus eliminating almost all background for object recognition.

For each of these feature spaces, we use a bag of visual words approach [13]. This approach make it possible to represent the object appearance as a fixed size occurrence histogram of features taken from a dictionary. These histograms are compared to histograms that have been memorized for different point of view of the learned objects through

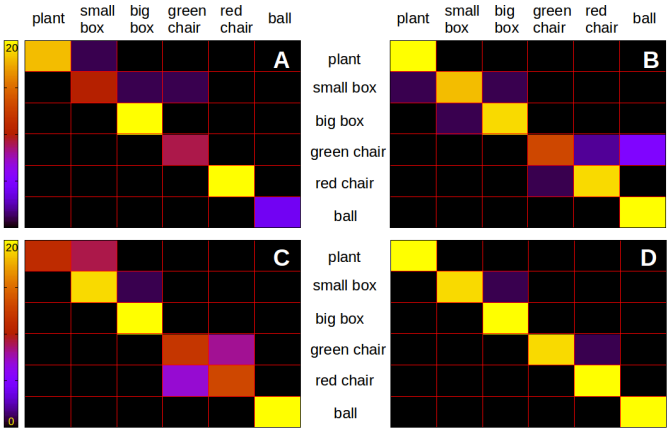

Fig. 5. Performance evaluation for object recognition using individual and combined methods. Figures A-C show results for color histograms, SURF and 3D surflet features. Figure D gives performance for the approach combining all three methods using a feed-forward neural network. Please note the frequent confusion of red and green chair objects for the SURF and surflet histograms methods (B and C), both of which are "colorblind". In contrast, the combined method (D) achieves disambiguation due to the inclusion of local color histograms.

a voting method [24]. As a result, we efficiently obtain a similarity score for the current object with all learned objects for each feature space individually (Fig. 4, Right). These similarity scores are then used as the input of a feed-forward neural network with one hidden layer (trained using back-propagation) that produces an overall similarity score as an output.

e) Experimental results: To rigidly evaluate our object recognition system, we created a small database of 6 objects seen from all viewpoints. We used 20 training views per object and 25 independently collected test view. Results for object recognition are given in Fig. 5. As can be clearly perceived, the recognition results strongly increase if the integration of several methods is used. Particularly instructive is the case of the red and green chairs: neither SURF nor surflet features have the possibility to discriminate identical chairs of different colours. Color histograms, however, can do this although their overall performance is inferior. The integration step combines the strength of each single method to give a much higher overall level of performance.

C. Ground and wall classification

The goal of this module is to classify the surfaces according to their type such as wood, carpet, gravel or concrete. This method is applied to the color image from the PTZ camera pointing to the floor just in front of the robot. It is also applied to regions of the image which the Kinect camera detected as walls. When images contain multiple surface types, only the type covering the largest area is estimated.

For this classification, we use the randomized trees approach proposed by [21]. This method is based on ensembles of extremely randomized decision trees that are used to predict the category associated to random sub-windows extracted from

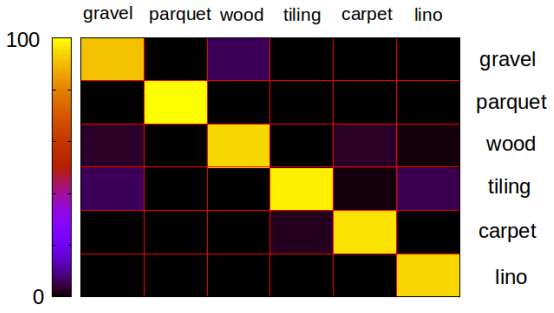

Fig. 6. Confusion matrix obtained for ground classification with 6 ground types. The error rates obtained for the ground types of gravel, parquet, wood, tiling, carpet and lino are, respectively, 9%,6%,2%,4% and 6%.

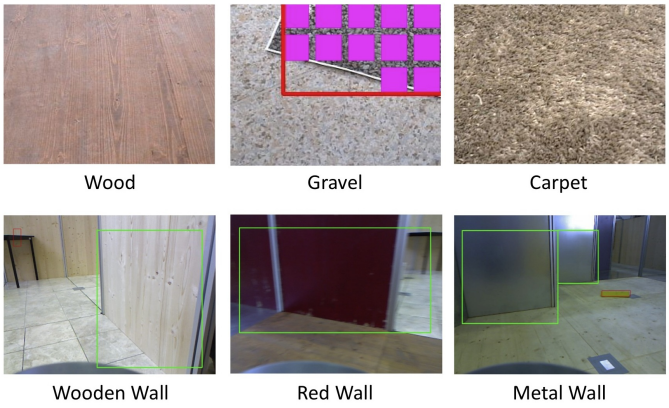

Fig. 7. Texture recognition examples. Top row: ground type recognized through PTZ camera. Bottom row : wall recognized through the Kinect camera. Wall bounding boxes (in green) have been constructed from 3D point clouds segmentation.

images. As an input for the method, we use image sub-windows of size 4x4 pixels represented as 48-dimensional vectors using the HSV color space. A majority vote using the categories associated with several sub-windows taken from a given image is then used to predict the image category. In order to filter out uncertain detections, the result of the classification is taken into account only if the difference in number of votes between the best and second best category is higher than a given threshold. This method requires few example images of each category for training. The use of the HSV color space makes it possible to robustly recognize colors independently of the illumination conditions by simply including color variations for each category in the training set.

Among the ground categories, gravel has to be treated differently because our robot is not able to roll on such a surface. As a consequence, gravel has to be localized more precisely in the image so as to be avoided. For this, we use the same classification algorithm, but apply it to regularly spaced 32x32 windows instead of the whole image. The bounding box of the windows containing gravel is projected onto the ground plane and transmitted to the obstacle avoidance module.

Figure 6 shows an example of the confusion matrix obtained with 6 ground types. The results are quite good and will

be further enhanced by subsequent filtering during semantic mapping. Figure 7 shows examples of the ground and wall categories used during the CAROTTE competition. The figure also illustrates the bounding box associated with gravel recognition and the bounding boxes of walls detected through the Kinect camera.

D. Multi-sensor obstacle detection

The environment proposed by the competition is artificial, but presents several difficulties for obstacles avoidance. In order to be robust to the presence of glasses, mirrors, 3D objects and gravel on the ground, we had to integrate multiple information sources to estimate the free space accessible to the robot.

The main navigation sensor in our system is the SICK laser which produces a 2D occupancy grid map containing most of the obstacles situated 30 cm above the ground. In order to detect smaller objects, glasses and mirrors that are not perceived by the laser, we use sonar information when it is not coherent with laser readings [25]. This information is added as an obstacle to the occupancy grid map. 3D objects like tables or benches under which the robot could get stuck are detected using the Kinect camera. For this, all the point clouds of the segmented objects up to the robot's height are projected onto the ground and their 2D convex envelope is added as obstacle to the occupancy grid map. Finally, the envelope of the gravel area detected using vision is also added as obstacle. The resulting safe space map is then used for path planning.

E. Semantic mapping

All the information produced by the previous modules is integrated to produce a semantic map. The 2D map is first segmented into rooms. For this, we detect the main orthogonal wall directions and search for doors, defined as openings of a given size along these walls. The connected components limited by walls and doors are assumed to be rooms, and the door position makes it possible to produce a topological map of the environment.

When multiple detections of a given object are encountered, the estimated positions are integrated using a Kalman Filter in order to produce precise position estimates. Each object is associated to the room in which it has been localized. The ground and wall types detected in each room are also integrated and only the categories with sufficient detections are kept for each room in order to filter out false detections. All this information is used to produce 2D annotated grid maps and an environment description as an XML file containing rooms, their connectivity, the objects they contain and the wall and ground types.

Figure 8 presents a map produced by our system in an 40 m^2 indoor environment that contains nine known objects (2 folding seat, drawers, 2 bottles, 1 chair, 1 red ball, 1 potted plant, 1 paper box). The top part shows the 2D map produced by the SICK laser, along with the obstacles detected by Kinect and by the gravel detection algorithm (grey polygons). Green and blue numbers indicate the positions of the detected ground and wall type. In this environment, only lino is present and ground type is correctly identified in all image except one, recognized as tiling. The wall categories are mostly correct, with some confusions between the 3 categories that are filtered out when integrating detections in the rooms. The bottom part shows the result of the room segmentation and the mean position of the detected objects inside each room. Seven objects out of nine present in the environment have been detected. The two objects have been missed because they do not appear completely in any image taken during exploration.

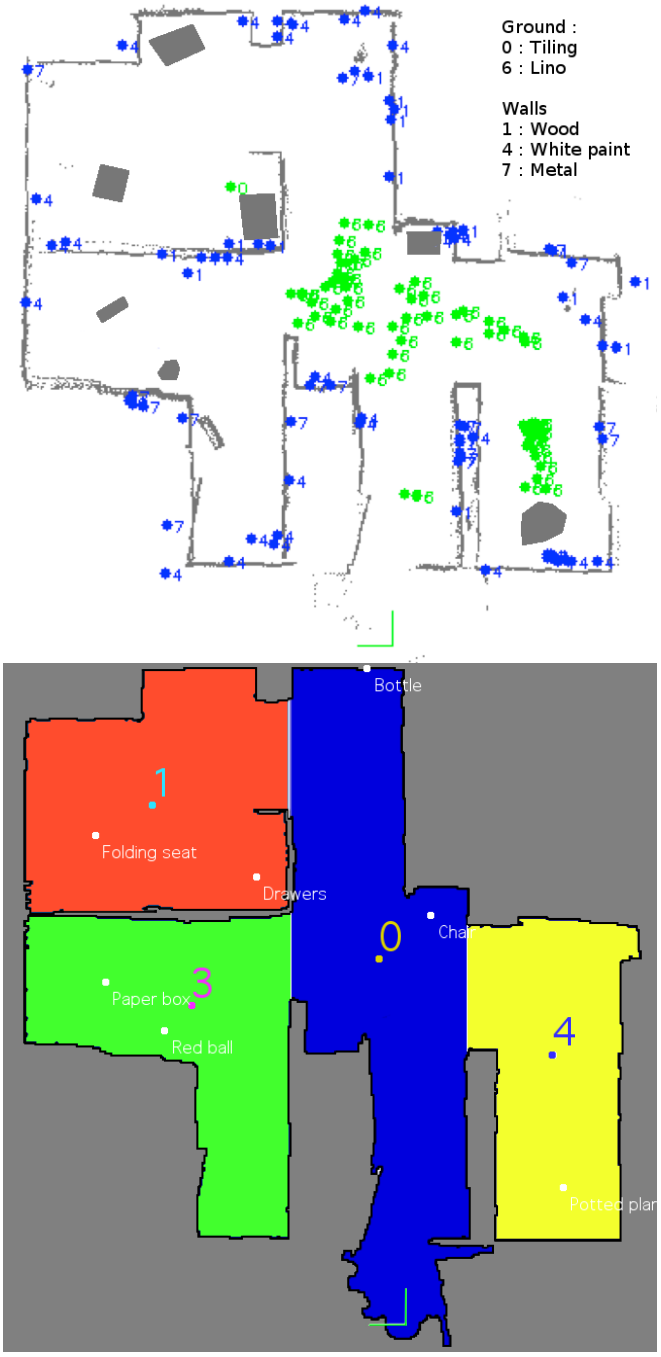

Fig. 8. Map produced by our system (see text for details).

The positions of the detected objects are within 20 cm of their true positions. The error is larger for bigger objects as it is more difficult to estimate a correct object position from a partial view of these objects.

IV. Conclusion and future work

Compared to our previous approach to object recognition using only color vision [4], the RGBD object recognition strongly improves the object recognition performance: an object that is correctly segmented using 3D information will be recognized with very high confidence. However, our current object segmentation approach is limited to objects that are isolated, and thus is more restrictive in this aspect than our previous vision-based approach. Overall, the system is however more reliable, producing less false alarms.

The new texture classification approach offers very good performances and produces very few erroneous results. The main limitation occurs for highly similar ground categories such as metal and concrete which are mostly uniformly grey and thus cannot be reliably distinguished. The wire fence wall type is also difficult to recognize as the environment behind is sometimes visually dominant in images. However, the global filtering in the final semantic mapping step is able to filter out most of these erroneous detections.

Finally, the obstacle avoidance strategy that integrates cues from multiple sensors has been very difficult to develop as it depends on most of the other system components. Long experimental testing has led to a system producing overall safe robot behaviour thanks to recovery strategies like the retracing of previous paths that will guide the robot in case of, e.g., erroneous gravel detections. However a more global and principled system approach would be necessary to ensure that such a core system component will behave correctly in all failure cases.

For the last CAROTTE competition that will take place in 2012, our future work will deal mainly with the improvement of the object segmentation module. In particular, we would like to be able to segment objects that are grouped together and objects that are put on shelves. We are also developing a 3D mapping method based on the RGBD camera in order to produce a visually appealing map and to use more global scene information for object segmentation.

Acknowledgment

The PACOM project is supported by DGA in the frame of the CAROTTE competition and funded by ANR under the subvention 2009 CORD 102.

References

[1] O. Martinez Mozos, R. Triebel, P. Jensfelt, A. Rottmann, and W. Burgard, "Supervised semantic labeling of places using information extracted from sensor data," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 391–402, 2007.

[2] D. F. Wolf and G. S. Sukhatme, "Semantic Mapping Using Mobile Robots," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 245–258, 2008.

[3] A. Aydemir, M. Göbelbecker, A. Pronobis, K. Sjöö, and P. Jensfelt, "Plan-based object search and exploration using semantic spatial knowledge in the real world," in *Proceedings of the 5th European Conference on Mobile Robots (ECMR'11)*, Örebro, Sweden, Sep. 2011.

[4] I. Jebari, S. Bazeille, E. Battesti, H. Tekaya, M. Klein, A. Tapus, D. Filliat, C. Meyer, S. Ieng, R. Benosman, E. Cizeron, J.-C. Mamanna, and B. Pothier, "Multi-sensor semantic mapping and exploration of indoor environments," in *Proceedings of the 3rd International Conference on Technologies for Practical Robot Applications (TePRA)*, 2011.

[5] M. Swain and D. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, 1991.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proceedings of the ECCV*, 2006.

[8] B. Schiele and J. L. Crowley, "Object recognition using multidimensional receptive field histograms," in *Proceedings of the European Conference on Computer Vision*, 1996.

[9] P. A. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[10] E. Wahl, U. Hillenbrand, and G. Hirzinger, "Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification," in *Proceedings of the Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM)*, 2003.

[11] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *Proc. of International Conference on Robotics and Automation (ICRA)*, 2010.

[12] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *CVPR*, 2010, pp. 998–1005.

[13] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proceedings of the International Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477.

[14] M. Riesenhuber and D. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, 1999.

[15] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *In ECCV workshop on statistical learning in computer vision*, 2004, pp. 17–32.

[16] A. Gepperth, J. Edelbrunner, and T. Bücher, "Real-time detection of cars in video sequences," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV2005)*, June 2005, pp. 625–631.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *In CVPR*, 2005, pp. 886–893.

[18] T. Leung and J. Malik, "Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons," *Int. J. Comput. Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[19] M. Varma and A. Zisserman, "A Statistical Approach to Texture Classification from Single Images," *Int. J. Comput. Vision*, vol. 62, no. 1-2, pp. 61–81, 2005.

[20] ——, "Texture classification: Are filter banks necessary?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[21] R. Marée, P. Geurts, J. Piater, and L. Wehenkel, "Random subwindows for robust image classification," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, C. Schmid, S. Soatto, and C. Tomasi, Eds., vol. 1. IEEE, June 2005, pp. 34–40.

[22] I. Jebari, S. Bazeille, and D. Filliat, "Combined vision and frontier-based exploration strategies for semantic mapping," in *Proceedings of the 3rd International Asia Conference on Informatics in Control, Automation and Robotics (CAR 2011)*, 2011.

[23] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL)," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[24] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2007.

[25] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *In Proceedings of the IEEE International Symposium on Computational Intelligence, Robotics and Automation*, 1997, pp. 146–151.